

RAFT: Adapting Language Model to Domain Specific RAG

Year: 2024 | Citations: 284 | Authors: Tianjun Zhang, Shishir G. Patil, Naman Jain

Abstract

Pretraining Large Language Models (LLMs) on large corpora of textual data is now a standard paradigm. When using these LLMs for many downstream applications, it is common to additionally bake in new knowledge (e.g., time-critical news, or private domain knowledge) into the pretrained model either through RAG-based-prompting, or fine-tuning. However, the optimal methodology for the model to gain such new knowledge remains an open question. In this paper, we present Retrieval Augmented FineTuning (RAFT), a training recipe that improves the model's ability to answer questions in a "open-book" in-domain settings. In RAFT, given a question, and a set of retrieved documents, we train the model to ignore those documents that don't help in answering the question, which we call, distractor documents. RAFT accomplishes this by citing verbatim the right sequence from the relevant document that would help answer the question. This coupled with RAFT's chain-of-thought-style response helps improve the model's ability to reason. In domain-specific RAG, RAFT consistently improves the model's performance across PubMed, HotpotQA, and Gorilla datasets, presenting a post-training recipe to improve pre-trained LLMs to in-domain RAG. RAFT's code and demo are open-sourced at github.com/ShishirPatil/gorilla.